Amendments to the Claims

Claim 1. (Currently Amended) A computerized storage and retrieval system of biological information comprising:

a means for data entry;

a means for displaying the data;

a programmable central processing unit for performing automated analysis; and

a data storage means containing protein pathways and annotated information on the pathways stored in a relational database, wherein the pathways are annotated and organized in a curated clustering arrangement and [~~wherein~~] the annotated information is accessed through the relational database.

Claim 2. (Currently Amended) The [~~computer~~] system of claim 1, wherein the information pertaining to the pathways is stored in a plurality of tables comprising proteins, their sequences and attributes; protein interactions; protein-protein associations; protein pathways; mRNA, microarray, and protein expression data; genes, their sequences and attributes; and descriptions of cells, tissues, organs, pathology reports, patient histories, and treatments.

Claim 3. (Currently Amended) The [~~computer~~] system of claim 1, wherein the central processing unit is programmed to retrieve, input, edit, annotate, search, calculate similarities, align, and predict homologous or orthologous protein pathways.

Claim 4. (Currently Amended) The [~~computer~~] system of claim 1, wherein the central processing unit is programmed to perform protein sequence analysis, protein interactions analysis, protein-protein association analysis, protein pathway analysis, gene expression analysis, pathway annotation analysis, pathway edit analysis, pathway expression analysis, tissue expression analysis, subtractive hybridization analysis, electronic northern analysis, or commonality analysis.

Claim 5. (Currently Amended) The [~~computer~~] system of claim 1, wherein the data is entered using the standard for pathway representation.

Claim 6. (Currently Amended) The [~~computer~~] system of claim 1, wherein a means for displaying the data is used to show two related pathways as a diagram containing nodes which represent proteins or non-protein molecules; modes that represent protein interactions or protein-protein associations; scores calculated from sequence, motif or structural homologies that interrelate nodes; and coefficients of similarity that interrelate modes of the pathway.

Claim 7. (Currently Amended) The [~~computer~~] system of claim 1, wherein the central processing unit is programmed to compare two protein pathways by a node-only, a mode-only, or a node-and-mode comparison and [~~wherein~~] the node-only comparison is selected from protein only, non-protein only, and protein and non-protein nodes.

Claim 8. (Currently Amended) The [~~computer~~] system of claim 1, wherein the central processing unit is programmed to run an optimization algorithm for dynamic programming comprising:

a) initializing an array, in which a two dimensional array M=M_{ij} with J rows and variant length for each row, the length for i-th row is n_i is set up and M_{Ji}=0, where 1<=i<=n_J,

b) backfilling the array via backward recursion with the formula

$$M_{ik} = \max_{\substack{j>i \\ 1\le l\le n_j}} \left\{ w\left(a_{ik}, a_{jl}\right) + M_{jl}\theta\left(w\left(a_{ik}, a_{jl}\right)\right)\right\} \text{ for } 1 \le k \le n_i,\ 1 \le i \le J$$

where θ(.) is the step function defined as θ (v)={0, if v<=0; 1, if v>0} and w(.,.) is the scoring function between the two nodes, defined as

$$w(a_{ik}, a_{jl}) = \begin{cases} 0, \text{ if i=j, } a_{ik}=a_{jl},\ a_{ik}=-D, \text{ or } a_{jl}=-D \\ \theta\left(c_{ik,jl}-t_c\right)\cdot\left\{\alpha\left(1-\left|s_{ik}-s_{jl}\right|\right)+(1-\alpha)c_{ik,jl}\right\} \text{ otherwise.} \end{cases}$$

where[~~and,~~] D>0

; and

c) using traceback to identify putative pathways PPW_j, 1<=j<=max n_i with the top *n* best scores.

Claims 9-22. (canceled).

Claim 23. (New) The system of claim 1, wherein the central processing unit is programmed to run a constrained clustering algorithm comprising:

a) assigning a distance between every pair of proteins in the database;

b) merging at each round the two closest pairs into a cluster until user-set threshold is met; and

c) computing mean linkage between two clusters using

$$d(c_1, c_2) = \left\{ \frac{1}{n_1 + n_2} \sum_{i\in c_1} \sum_{j\in c_2} d_{ij}{}^{p} \right\}^{1/p}$$

wherein n_1 is the size of c_1 and n_2 is the size of c_2 and the distance between two clusters is a weighted average of the distance between two proteins i and j and a function of similarity to query protein ($s_i=S_{iq(i)}$, $s_j= S_{jq(j)}$)

where q(i) is the query protein homologous to database protein i and protein-protein association (a_{ij}) and

$$d_{ij} = \alpha\left(1-a_{ij}\right)+\left(1-\alpha\right)\left|s_i - s_j\right| .$$

Claim 24. (New) The system of claim 3, wherein genes that encode known proteins are used to annotate modes of a pathway comprising:

a) selecting genes which encode known proteins;

b) employing the genes to produce a protein-protein association matrix containing coefficients of similarity; and

c) using the coefficients of similarity from the matrix to annotate the modes of the pathway.

Claim 25. (New) The system of claim 7, wherein pathway analysis uses a node-and-mode comparison comprising:

a) submitting a query pathway and protein sequences;

b) comparing nodes using an optimization algorithm for dynamic programming wherein a sequence identity score or p-value summarizes similarity and a weighting factor between 0 and 1 is assigned to corresponding nodes;

c) comparing modes by generating a SCIM matrix wherein a coefficient of similarity is assigned to corresponding modes;

d) aligning pathways globally or locally, wherein insertion or deletion of nodes or modes incurs a penalty;

e) summing all similarity scores; and

f) displaying at least one high-scoring segment of the aligned pathways derived from node-and-mode comparison.

Claim 26. (New) The system of claim 1, wherein the central processing unit is programmed to perform pathways analysis using a submitted query pathway and protein sequences comprising:

a) organizing the pathway and sequences using the standard for pathway representation;

b) comparing protein sequences of the query pathway with all protein sequences in the pathways database using standard methods of protein comparison;

c) using a SCIM matrix to compare coefficients of similarity for each interaction of the query pathway and all interactions for proteins in the pathways database;

d) calculating an OS-score based on sequence identity and coefficients of similarity;

e) removing all pathways not meeting a user-specified threshold for OS-score; and

f) retrieving aligned pathways meeting the threshold, thereby performing pathways analysis.

Claim 27. (New) The system of claim 3, wherein pathways database is searched for protein interactions comprising:

a) submitting a query pathway;

b) performing protein interactions analysis between the query pathway and all pathways in the pathways database wherein coefficient of similarity is produced to interrelate each mode of the query pathway and a mode of the most closely related protein pathway;

c) retrieving at least one pathway alignment; and

d) showing the alignment based on the search for protein interactions.

Claim 28. (New) The system of claim 3, wherein a query pathway is used to search a pathways database to predict homologous pathways comprising:

a) submitting a query pathway and protein sequences;

b) comparing the query pathway and protein sequences with all pathways and proteins in the pathways database; and

c) retrieving a plurality of pathway alignments based on OS-score, thereby predicting homologous pathways.

Claim 29. (New) The system of claim 3, wherein a query pathway to search a pathways database to predict orthologous pathways comprising:

a) submitting a query pathway and known protein sequences;

b) comparing known sequences to all protein sequences stored in the database;

c) retrieving orthologous proteins with the highest identity to the known proteins;

d) inheriting protein interactions from the query pathway; and

e) aligning the query pathway and the orthologous proteins, thereby predicting orthologous pathways.

Claim 30. (New) The system of claim 7, wherein a known protein pathway is used to predict the nodes and modes of a novel pathway comprising:

a) submitting a known protein pathway and its protein sequences;

b) applying standard methods of comparison to determine similarity between the known protein sequences and protein sequences in the databases, thereby predicting candidate nodes;

c) utilizing coefficients of similarity from protein interactions or protein-protein association data to predict candidate modes; and

d) retreiving a pathway with an OP-score obtained using an optimization algorithm, thereby predicting the nodes and modes of a novel pathway.

Claim 31. (New) The system of claim 30, wherein coefficients of similarity are selected from RNA/cDNA counting, microarray expression, protein expression, known protein-protein associations, a promoter similarity matrix, and more than one of these methods.

Claim 32. (New) The system of claim 30, wherein the constrained clustering method is selected from average linkage, single linkage, complete linkage, K-means, or self-organizing maps with the constraint that no more than one protein in each cluster is derived from a single column of aligned proteins and the accuracy of the prediction is determined by an OP-score.

Claim 33. (New) The system of claim 3, wherein novel pathways are predicted comprising:

a) generating candidate proteins from at least one species for each node based on a protein search;

b) employing an optimization means to find likely linear linkages between candidate proteins aligned to the query pathway with possible gaps in the alignment; and

c) reporting all pathways having optimal and sub-optimal predictions that satisfy user-specified alignment and interaction parameters wherein the accuracy of the prediction is provided by OP-score.

Claim 34. (New) The system of claim 33, wherein the optimization means is selected from linear next-neighbor criteria, global minimization criteria, optimization algorithm for dynamic programming, and iterative searches using at least two of these means.

Claim 35. (New) The system of claim 1, wherein the central processing unit is programmed to determine the function of a protein or a gene that encodes the protein comprising:

a) placing the protein encoded by the gene in a candidate pathway involving at least two proteins; and

b) using protein interactions with proteins and non-protein molecules, cellular location, and expression, thereby determining the function of the protein or the gene encoding the protein.

Claim 36. (New) The system of claim 1, wherein the central processing unit is programmed to predict novel pathways comprising:

a) submitting a query pathway and protein sequences;

b) processing the query pathway and protein sequences using orthologous pathway prediction wherein the data is derived from protein similarities and interactions, homologous pathway prediction wherein the data is derived from protein similarities and interactions, and protein-protein associations; and

c) applying an optimization algorithm for dynamic programming or a constrained clustering algorithm, thereby predicting novel pathways.